GREENEYES: NETWORKED ENERGY-AWARE VISUAL ANALYSIS

L. Baroffio[1], *M. Cesana*[1], *A. Redondi*[1], *M. Tagliasacchi*[1],
J. Ascenso[2], *P. Monteiro*[2], *E. Eriksson*[3], *G. Dán*[3], *V. Fodor*[3]

[1]DEIB, Politecnico di Milano - Italy
[2]KTH, Royal Institute of Technology - Stockholm, Sweden
[3]Instituto Superior Técnico, Instituto de Telecomunicações - Lisbon, Portugal

ABSTRACT

The GreenEyes project aims at developing a comprehensive set of new methodologies, practical algorithms and protocols, to empower wireless sensor networks with vision capabilities. The key tenet of this research is that most visual analysis tasks can be carried out based on a succinct representation of the image, which entails both global and local features, while it disregards the underlying pixel-level representation. Specifically, GreenEyes will pursue the following goals: i) energy-constrained extraction of visual features; ii) rate-efficiency modelling and coding of visual feature; iii) networking streams of visual features. This will have a significant impact on several scenarios including, e.g., smart cities and environmental monitoring.

Index Terms— Visual Sensor Networks, Local Visual Features, Visual Analysis

1. INTRODUCTION

Reading a book, recognizing a familiar face, etc. are all actions that characterize peoples everyday life and require processing of visual stimuli. Such stimuli are analyzed and ultimately converted into high level semantic concepts by the human early visual system, a process whose metabolic energy expenditure is extremely low.

Digital smart cameras have been developed mimicking a simplified model of the human visual system. Images, or image sequences, are acquired in digital format, compressed in order to be stored and/or transmitted and finally analyzed to accomplish high level tasks, e.g. recognizing letters, faces, objects, detecting events, etc. Such a Compress-then-Analyze (CTA) paradigm is being successfully employed in a number of application scenarios, such as video surveillance. However, it is partially at odds with the paradigm adopted by the human visual system: first, image analysis is often based on a compressed and hence lossy representation of the original image, which might significantly impair its efficiency. Second, image data are stored and transmitted retaining a pixel-level representation, although only their semantics might ultimately matter. Finally, energy-efficiency is often regarded as a secondary aspect, or is neglected altogether, since most of the processing burden associated to common image analysis tasks is to be carried out at a centralized, power-eager computational node.

Conversely, the potential of the Internet of Things is leading to an ambitious long-term vision, in which battery-operated sensing nodes equipped with cameras and capable of wireless communication are organized in a so-called Visual Sensor Network (VSN). VSNs will play a major role in the Internet of Thing paradigm, enabling several novel applications such as object recognition, event detection, localization and tracking. This vision requires departing from the traditional Compress-then-Analyze solution and pursuing a paradigm shift that affects the way visual data is sensed, processed and transmitted. The key tenet is that most visual analysis tasks can be carried out based on a succinct representation of the image, which entails both global and local features, while it disregards the underlying pixel-level representation. Consequently, the GreenEyes project proposes the adoption of a novel Analyze-then-Compress (ATC) paradigm, in which camera nodes extract visual features from the acquired content and transmit a compressed version of such features to a remote device in order to be analyzed. To this end, GreenEyes is working towards the following goals:

1. *Energy-constrained extraction of visual features:* computing visual features in battery-operated nodes with an embedded microprocessor and radio chip on board is extremely challenging, thus calling for new computational primitives and practical algorithms specifically tailored to such kind of hardware.

2. *Rate-accuracy modeling and coding of visual features:* GreenEyes will take a fresh perspective, by proposing novel joint feature extraction and coding schemes, which are specifically designed to fulfill the computation, coding and efficiency requirements of VSNs

3. *Networking streams of visual features:* GreenEyes will define a set of new protocols and methodologies to: (i) coordinate distributed features extraction; (ii) develop inter-

camera coding of visual features; (iii) provide efficient delivery of visual features over multi-hop VSNs.

2. FEATURE EXTRACTION

The GreenEyes project aims to adapt feature extraction algorithms to the context of resource-constrained architectures and to introduce novel techniques and algorithms that allow the extraction of high-quality features at a low computational cost. The main achieved results are listed as follows:

- **Optimized extraction of binary local features for ARM architectures:** Many of todays smartphones and VSN platforms are based on ARM architectures. In [1], we propose an optimized version of the binary features extraction algorithm BRISK, resorting to solutions tailored to the peculiar nature of the computing architectures. With the proposed modifications, the computational time is reduced by as much as 30%, without affecting the quality of the generated features.
- **Efficient detection of local features using temporal redundancy:** Several visual analysis applications such as object tracking, event detection, surveillance, require visual features to be extracted on a frame-by-frame basis. In this context, we propose a fast keypoint detection architecture tailored to the context of video sequences, which exploits temporal redundancy between acquired frames to speed up the detection process [2]. Our experiments show that it is possible to achieve a reduction in computational time by up to 40%.
- **Binary descriptors from asymmetric boosting (BAMBOO):** Traditional feature extraction algorithms are mostly handcrafted, and designed according to intuitions coming from the observation of common visual patterns. Instead, we use machine learning to design an algorithm able to extract semantic information from visual content. The proposed BAMBOO descriptor [3] achieves better accuracy performance compared to other state-of-the-art feature extraction algorithms, at a similar computational cost.
- **Local Binary Feature Selection for Compact Representation:** Feature detection algorithms usually let the user select the sensitivity of the extraction algorithm, leading to a tradeoff between the number of features extracted from a frame and their quality. In this context, we propose a novel algorithm that assigns a score to each detected feature capturing the importance of the feature to achieve good matching performance [4]. Experiments show that carefully selecting the features to be transmitted allows to achieve from 35% to 60% of bitrate reduction for the same target accuracy.

3. FEATURE CODING

GreenEyes proposes several coding schemes for local visual features, which are specifically designed to fulfill the computation, coding and efficiency requirements of VSNs. These techniques can be broadly classified as:

1. *Intra-descriptor coding descriptor schemes:* each descriptor is independently encoded from the others. In this case, only the correlation between the descriptor elements (dexels) is exploited, typically considering their importance with respect to the accuracy of the visual analysis task.

2. *Inter-descriptor coding descriptor schemes:* each descriptor is encoded with respect to other reference descriptors obtained from the same image, from previous decoded images or even from previous decoded views (in a multi-view scenario).

Considering that the the computation of low bitrate descriptors is rather important for several applications that are limited by bandwidth, storage or latency, the proposed techniques exploited in GreenEyes are:

- **Coding of real-valued local descriptors:** A study and comparison of the rate-distortion efficiency of different lossy coding solutions such as quantization, Karhunen-Loeve transform, sparse coding and open-loop (tree) coding was conducted. The proposed techniques target the correlation within the dexels and among descriptors extracted from different keypoints of the same image [5]. In addition, a coding architecture designed for local features extracted from video sequences was developed, in this case exploiting both spatial and temporal redundancy using novel motion estimation and mode decision algorithms [6].
- **Coding of binary local descriptors:** Binary descriptors are a promising representation that provides an alternative to real-valued descriptors with high efficiency and low complexity. To obtain a more compact representation of binary descriptors, lossless methods such as descriptor sorting, predictive coding and adaptive binary arithmetic entropy coding were proposed in [7]. In [8], two clustering algorithms that group similar descriptors and find an efficient prediction path between them are proposed. Also, the space of binary descriptors can be divided into clusters, and each cluster can be represented with a centroid that can be used for prediction. To further obtain higher compression ratios at the cost of a slightly lower performance, GreenEyes developed a lossy coding scheme that selects the most discriminant dexels and discards the remaining ones [9]. A coding architecture suitable for binary local features extracted from video content was also investigated in [10] and suitable coding mode decision algorithms to find the best descriptor coding modes were also proposed in [11].

- **Multi-View coding of binary descriptors:** Visual content can also be acquired and analyzed directly in spatially distributed sensing nodes, which can cooperate to perform efficient coding of local descriptors. GreenEyes has also investigated two possible solutions with different complexity-efficiency tradeoffs: i) inter-node predictive coding schemes, where each sensing node encodes visual features based on previously received features from other neighboring nodes; ii) distributed source coding of visual features where features are independently encoded but jointly decoded to reduce the rate necessary for efficient image analysis; this last solution allows to achieve lower complexity at the sensing node.

Nowadays, scalable binary descriptor representations are also being investigated to cope with wireless channels where bandwidth resources can change over time providing fine rate adaptation to the available node and network resources. In such a case, a set of hierarchical layers is constructed to achieve the best performance for a certain bitrate budget, namely by selecting the most discriminative descriptors [9] and descriptor elements [4]. Also, error resilience and error concealment techniques were proposed to achieve graceful degradation for noisy channels.

4. FEATURE NETWORKING

Besides feature extraction and coding, GreenEyes aims to develop algorithms and protocols for efficient and timely feature processing and transmission in a network of energy constrained devices. The main results are as follows:

- **Workload characterization:** Efficient distributed processing of visual features requires an understanding of the spatial and the temporal characteristics of streams of features. In [12] we characterize the density distribution, the spatial distribution and the correlation of interest point locations in a large image data set, and analyze the potential gains of various computation off-loading schemes for distributed visual feature processing.
- **Performance modeling:** Timely processing of visual features in a network of nodes requires controlling the workload put on the system, and at the same time it requires proper dimensioning of the communication and computational resources. In [13] we proposed a low-complexity predictor to control the computational load of extracting features from a video sequence. In [14] we developed an analytical model of the completion time of networked visual feature processing as a function of the wireless capacity and the available computational resources based on stochastic network calculus.
- **Distributed extraction of features:** When distributing the processing of visual features at several sensor nodes, the dynamics of the visual content and the randomness of the wireless channels make the optimization of the allocation of processing tasks challenging. In [15] we provided closed form expressions for the minimization of the completion time of distributed feature extraction of a single image. In [16] we proposed an algorithm for minimizing the completion time for a sequence of images based on stochastic optimization and showed that the dynamics of the visual content have a larger impact on the completion time that the dynamics of the wireless channel. In [17] we proposed algorithms for distributed feature processing for several visual sensors that share a pool of processing nodes, and showed that algorithms that require little coordination may perform reasonably well.
- **Resource management:** Performing visual analysis tasks under delay and energy constraints in visual sensor networks requires the coordination of processing. In [18] we considered the problem of distributing BoW in a sensor network to accelerate recognition tasks and compared the performance of distributed recognition to that of a centralized solution.

5. APPLICATION SCENARIOS

The GreenEyes project targets all those applications where one or multiple cameras are used to perform advanced analysis tasks by transmitting the acquired visual content to a remote server through wireless communication. We focus in particular on those cases where camera nodes are characterized by tight bandwidth and energy constraints, and we analyze the benefits of adopting the newly proposed Analyze-then-Compress paradigm compared to the traditionally used Compress-then-Analyze paradigm. In the following, we list several application scenarios in which GreenEyes project results might have a significant impact:

- **Environmental Monitoring:** Networks of battery-operated and low-cost wireless cameras may be easily deployed to support applications related to environmental monitoring (e.g., recognizing animals in natural environments, detecting fires, land slides or other hazardous situations in remote or inaccessible area). In such scenarios, bandwidth availability is limited, considering that low-rate and low-power wireless technologies are generally employed to provide communication capabilities. In this context, we have shown that at low bitrates the ATC approach is not only the preferable solution, but also the only one that can be adopted [6].
- **Mobile Visual Search:** With the proliferation of handheld devices capable of acquiring images and video from the environment, Mobile Visual Search has gained increasing popularity in the past few years. It requires transmitting visual data from the user terminal to a remote server, which analyses it and replies back with information on the acquired content. The very same setup is also used in other two application scenarios similar to Mobile Visual Search,

such as Augmented Reality and Camera Based Indoor Positioning. Again, we have shown that the choice of the paradigm (CTA or ATC) is dictated by the bandwidth requirements. At low bitrate, ATC is the preferable solution [6].

- **Automotive:** Nowadays cars come equipped with several cameras for increasing the safety of drivers and passengers. The visual streams from each camera are typically transmitted to a wired central processor, which analyses the data and performs several tasks such as pedestrian and lane detection. The use of a wired technology allows for very high bandwidth availability. However, since the number of cameras may easily exceed the dozen, transmitting compressed videos from each camera to the central processor may be unfeasible. Therefore, the use of an ATC-like approach, where cameras perform features extraction and compression in a distributed manner and transmit the features to the central processor, may be beneficial for applications in the automotive field.
- **Privacy-aware applications:** Privacy awareness is a crucial aspect of every application where cameras acquire, transmit and store visual data that contains people. In the field of visual surveillance, a lot of attention has been recently given to those techniques that are able to analyze visual data in a privacy-aware manner, inferring high-level visual semantics while not revealing the identity of people. In such a context, and regardless of the bandwidth availability, it may be preferable to adopt the ATC paradigm instead of CTA: as a matter of fact, the transmission, storage and manipulation of features-based data instead of pixel-domain data constitutes a natural way of ensuring privacy.

6. CONCLUSIONS

The GreenEyes project aims at developing a comprehensive set of new methodologies, practical algorithms and protocols, to empower wireless sensor networks with vision capabilities. The key tenet of this research is that most visual analysis tasks can be carried out based on a succinct representation of the image, which entails both global and local features, while it disregards the underlying pixel-level representation. The project will have a significant impact on several scenarios, including smart cities and environmental monitoring.

7. REFERENCES

[1] L. Baroffio, A. Canclini, M. Cesana, A. Redondi, and M. Tagliasacchi, "Briskola: Brisk optimized for low-power arm architectures," in *Image Processing (ICIP), IEEE Intl. Conf. on*, Oct 2014, pp. 5691–5695.

[2] L. Baroffio, M. Cesana, A. Redondi, and M. Tagliasacchi, "Fast keypoint detection in video sequences," in *submitted to IEEE Intl. Conf. on Image Processing (ICIP)*, 2015.

[3] L. Baroffio, M. Cesana, A. Redondi, and M. Tagliasacchi, "BAMBOO: A fast descriptor based on asymmetric pairwise boosting," in *Image Processing (ICIP), IEEE Intl. Conf. on*, Oct 2014.

[4] P. Monteiro, J. Ascenso, and F. Pereira, "Local feature selection for efficient binary descriptor coding," in *Image Processing (ICIP), IEEE Intl. Conf. on*, Oct 2014, pp. 4027–4031.

[5] A. Redondi, M. Cesana, and M. Tagliasacchi, "Low bitrate coding schemes for local image descriptors," in *Intl. Workshop on Multimedia Signal Processing*, sept. 2012, pp. 124 –129.

[6] L. Baroffio, M. Cesana, A. Redondi, M. Tagliasacchi, and S. Tubaro, "Coding visual features extracted from video sequences," *Image Processing, IEEE Transactions on*, vol. 23, no. 5, pp. 2262–2276, May 2014.

[7] J. Ascenso and F. Pereira, "Lossless compression of binary image descriptors for visual sensor networks," in *Digital Signal Processing (DSP), 2013 18th Intl. Conf. on*, July 2013, pp. 1–8.

[8] P. Monteiro and J. Ascenso, "Clustering based binary descriptor coding for efficient transmission in visual sensor networks," in *Picture Coding Symposium (PCS)*, Dec 2013, pp. 25–28.

[9] A. Redondi, L. Baroffio, J. Ascenso, M. Cesana, and M. Tagliasacchi, "Rate-accuracy optimization of binary descriptors," in *Image Processing (ICIP), 20th IEEE Intl. Conf. on*, Sept 2013, pp. 2910–2914.

[10] L. Baroffio, J. Ascenso, M. Cesana, A. Redondi, and M. Tagliasacchi, "Coding binary local features extracted from video sequences," in *Image Processing (ICIP), 2014 IEEE Intl. Conf. on*, Oct 2014, pp. 2794–2798.

[11] P. Monteiro and J. Ascenso, "Coding mode decision algorithm for binary descriptor coding," in *Signal Processing Conf. (EUSIPCO), Proceedings of the 22nd European*, Sept 2014, pp. 541–545.

[12] G. Dán, M. Khan, and V. Fodor, "Characterization of SURF and BRISK interest point distribution for distributed feature extraction in visual sensor networks," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 591–602, 2015.

[13] E. Eriksson, G. Dán, and V. Fodor, "Prediction-based load control and balancing for feature extraction in visual sensor networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE Intl. Conf. on*, May 2014, pp. 674–678.

[14] H. Al-Zubaidy, G. Dán, and V. Fodor, "Performance of in-network processing for visual analysis in wireless sensor networks," in *accepted to IFIP Networking*, 2015.

[15] A. Redondi, M. Cesana, M. Tagliasacchi, I. Filippini, G. Dán, and V. Fodor, "Cooperative image analysis in visual sensor networks," *Ad Hoc Networks*, vol. 28, no. 0, pp. 38 – 51, 2015.

[16] E. Eriksson, G. Dán, and V. Fodor, "Real-time distributed visual feature extraction from video in sensor networks," in *Distributed Computing in Sensor Systems (DCOSS), 2014 IEEE Intl. Conf. on*, May 2014, pp. 152–161.

[17] E. Eriksson, G. Dán, and V. Fodor, "Algorithms for distributed feature extraction in multi-camera visual sensor networks," in *accepted to IFIP Networking*, 2015.

[18] S. Paris, A. Redondi, M. Cesana, and M. Tagliasacchi, "Distributed object recognition in visual sensor networks," in *Communication (ICC), 2015 IEEE Intl. Conf. on*, 2015.